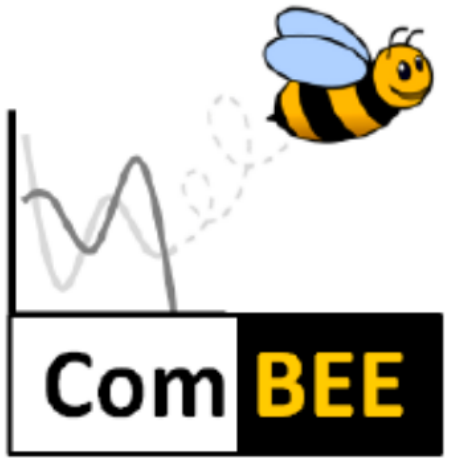
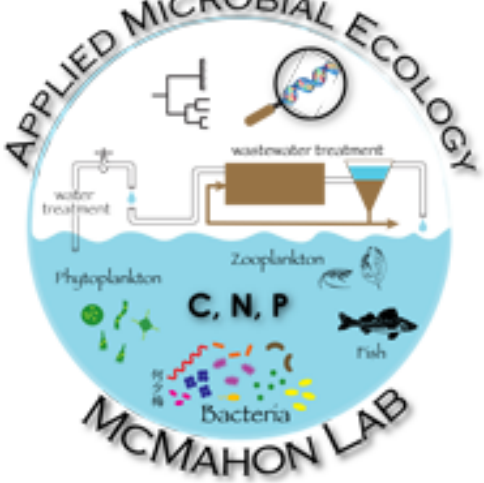
# ComBEE ‘Omics Study Group: Genome/Metagenome Assembly

Benjamin Peterson



# Genome/metagenome assembly

- What is it?
- Why should you do it?

# Six steps to a good assembly

1. Experimental Design/DNA sequencing

# Six steps to a good assembly

1. Experimental Design/DNA sequencing
2. Processing and QC on raw sequencing reads

# Six steps to a good assembly

1. Experimental Design/DNA sequencing
2. Processing and QC on raw sequencing reads
3. Grouping (or not) of metagenomes

# Six steps to a good assembly

1. Experimental Design/DNA sequencing
2. Processing and QC on raw sequencing reads
3. Grouping (or not) of metagenomes
4. Assembly itself!

# Six steps to a good assembly

1. Experimental Design/DNA sequencing
2. Processing and QC on raw sequencing reads
3. Grouping (or not) of metagenomes
4. Assembly itself!
5. Post-assembly processing (”fiddling”)

# Six steps to a good assembly

1. Experimental Design/DNA sequencing
2. Processing and QC on raw sequencing reads
3. Grouping (or not) of metagenomes
4. Assembly itself!
5. Post-assembly processing (”fiddling”)
6. Denial

# Step 1: Experimental Design/DNA sequencing

1. First need to decide on an appropriate strategy

    a. What's your question?
    b. What do you need to answer that question?
        a. Bins? Strain variants? Rare genes?

2. Next need to pick an approach

    a. Long vs. short reads/sequencing platform
    b. Number of samples
    c. Depth of sampling

# Step 1: Experimental Design/DNA sequencing

3. Prepare for and perform sequencing

    a. Considerations for library prep:
        a. Short reads – DNA quality, insert size
        b. Long reads – DNA quality and quantity, degree of shearing

# Step 1: Experimental Design/DNA sequencing

1. My needs
    1. In each of my systems, I'm interested in one particular gene (*hgcA*) and in the organisms that contain it.
    2. We didn't know how abundant they are.
    3. We also want to be able to reconstruct high quality genomes
    4. Often low quantities of DNA

2. My approach

    1. Focus on using very deep sequencing with short reads (Illumina HiSeq/NovaSeq)
    2. Try to obtain high quality DNA for future long read sequencing

# Step 2: Quality control

1. Inspect data (using FastQC)
2. Removing adaptors (if needed)
3. Digital normalization (remove redundant reads)
    1. (I haven’t done this)
4. Quality trim the raw reads
    1. Trimmomatic
    2. Sickle
5. Final inspection with FastQC

# Step 2: Quality control

1. My approach
    1. Inspect data (using FastQC)
    2. Quality trim the raw reads using Sickle
        1. Phred score cutoff of 20
        2. Minimum read length 100bp
    3. Re-use FastQC to check on the trimmed reads

# Step 3: Grouping (or not) of metagenomes

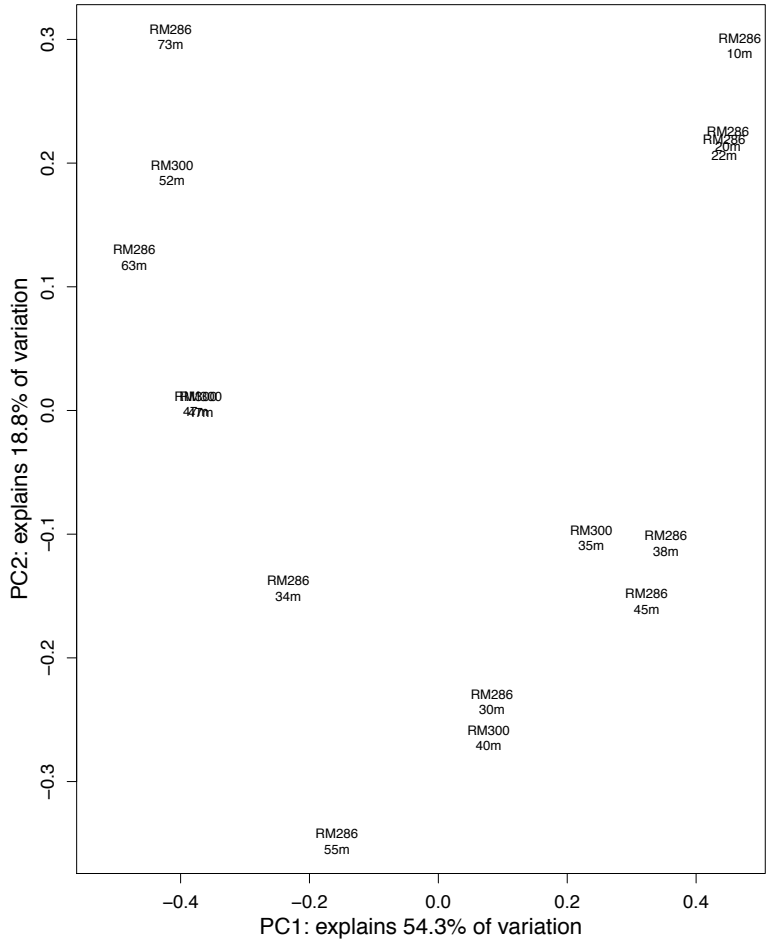
1. Advantages of coassemblies
    1. Fewer sets of genes to work with
    2. Effectively deeper sequencing can lead to better assemblies
2. Drawbacks of coassemblies
    1. Too deep of sequencing can lead to fragmentation due to sequencing errors
    2. Different strains within an assembly can also lead to fragmentation
    3. Longer assembly time and more memory-intensive

# Step 3: Grouping (or not) of metagenomes

1. My situation and needs
    1. 14 metagenomes
    2. Multiple redox states of interest
    3. I wanted to find as many potential hgcA genes as possible
    4. Access to computer power and space
    5. Ancillary water chemistry data
    6. Preliminary 16S sequencing information

# Step 3: Grouping (or not) of metagenomes

1. My approach
    1. Multiple co-assemblies of clusters of metagenomes
        1. Clusters determined separately by redox status and 16S composition, then combined
    2. Also ran a full coassembly to cover my bases



# Step 4: Actually assemble it

1. 2 types of assemblers
    1. De Bruijn graph
    2. Overlap-Layout-Consensus
    3. Alignments are used to manually clean assemblies or in hybrid assemblers

2. Comparison of assemblers:
    1. Short read, for metagenomes – DOI:10.1371/journal.pone.0169662
    2. Long read, for genomes - https://github.com/rrwick/Long-read-assembler-comparison

# Step 4: Actually assemble it

1. Short-read assemblers (for metagenomes)
    1. Mostly de Bruijn graph assemblers with multiple kmers
        1. metaSPADes seems to be most commonly used, but is memory intensive
        2. MEGAHIT as a faster option
    2. Key is to take advantage of the multiple kmers
2. My approach:
    1. Initially ran each assembly with 3-4 different assemblers to compare
    2. Now use metaSPADes to assemble smaller clusters of metagenomes
    3. MegaHit is used if the coassemblies include enough reads
        1. Does also seem to perform better in sediments

# Step 4: Actually assemble it

1. Long-read assemblers
    1. Most long-read assemblers are designed for genome assembly
    2. metaFlye recently developed as a stand-alone *de novo* assembler

2. Hybrid assemblers
    1. Canu – Not designed for metagenomes, but used for it
    2. hybridSPADes – reads in long and short reads, from any source
    3. OPERA-MS – new assembler designed specifically hybrid metagenomic assemblies

# Step 5: Post-assembly processing (“fiddling”)

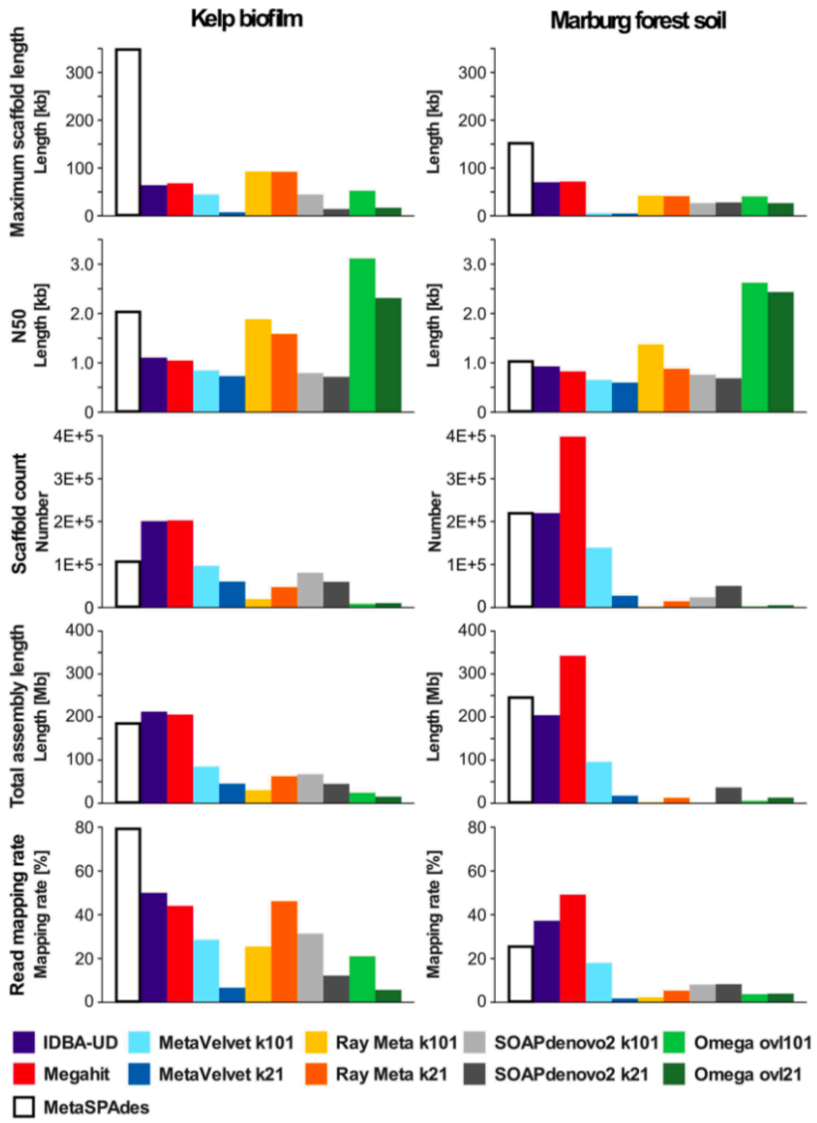
Beware ye who enter here…

Here there be rabbit-holes

# Step 5: Post-assembly processing ("fiddling")

First, check quality of assembly

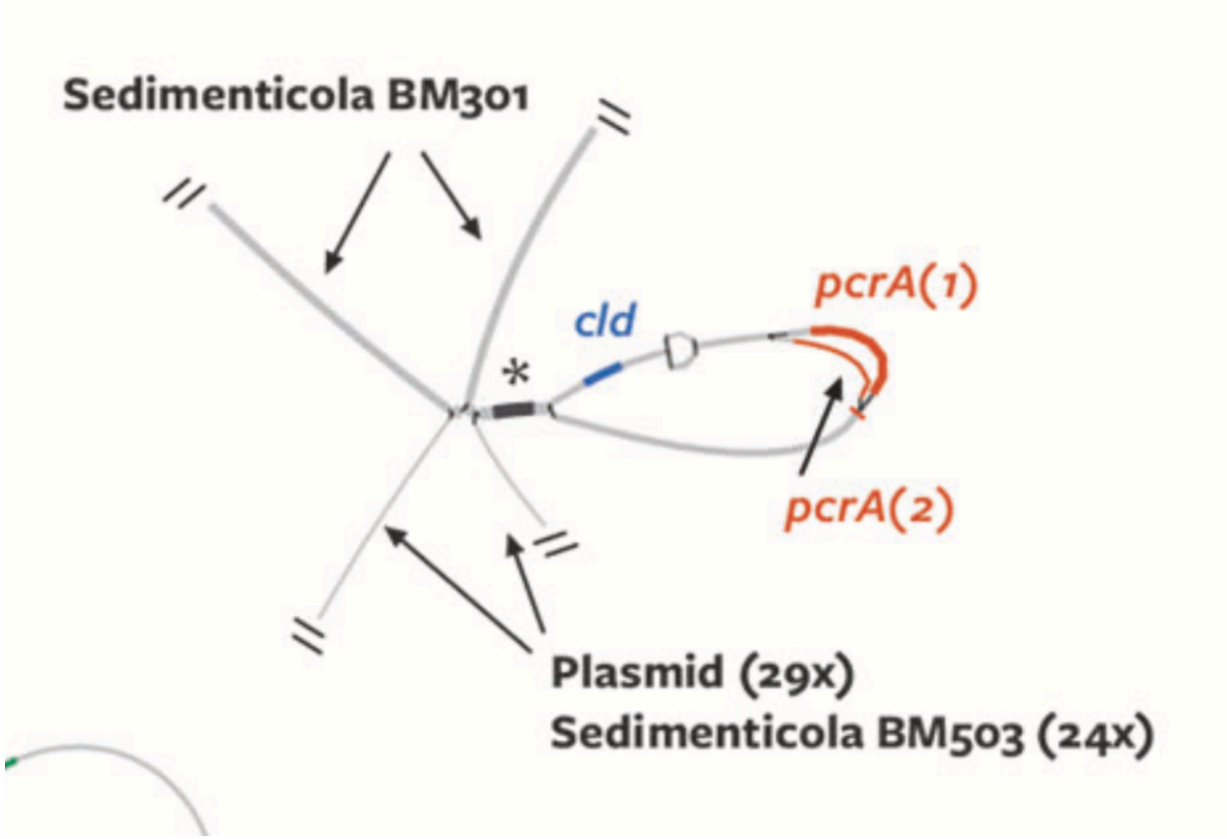
1. Number of scaffolds
2. Length of assembly
3. Fraction of reads mapping to assembly
4. Number of predicted genes
5. Average length of scaffold
6. N50 – 50% of the bases in the assembly are in contigs at least this long
7. L50 – smallest number of contigs that could contain up to 50% of all bases in an assembly



# Step 5: Post-assembly processing (“fiddling”)

Manual inspection of assembly

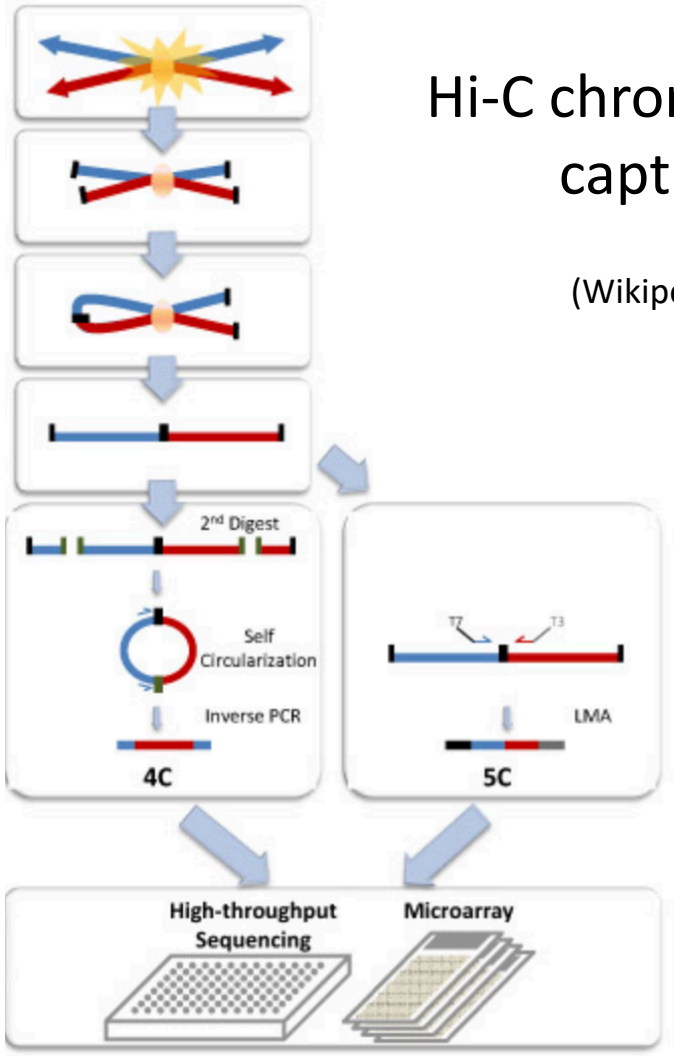
1. Bandage – program to visualize and search *de Bruijn* assembly graphs
    1. Can find evidence for laterally transferred contigs this way
2. Geneious – Easy visualization of contigs and mapping



# Step 5: Post-assembly processing (“fiddling”)

Assembly improvements

1. Hi-C-informed metagenome deconvolution (doi:10.1038/s41467-018-03317-6)
2. Pilon: error correction of long reads/contigs using short Illumina reads
    1. Many more like this
3. Reassembly using subset of reads




Hi-C chromosome capture

(Wikipedia)

# Step 5: Post-assembly processing (“fiddling”)

1. My approach
    1. Reassembled with randomly subsetted reads to try to assemble super abundant genes and their neighborhoods
    2. Checked for transfer of sequences with Bandage
    3. Reassembly of binned sequences using metaSPADes or Geneious
    4. And many more…

# Step 6: Denial


The final question isn’t whether or not it’s the best it can possibly be.

The final question is, can you sufficiently answer the questions you want to with the quality of the data you have?